Estimation of network leakage with smart meters

An analysis of sampling requirements

Prepared for
The Great Britain gas distribution networks

9 August 2016

www.oxera.com

oxera
compelling economics

Contents

Executive summary 1

1 Introduction and overview 4

1.1 Background 4
1.2 Scope of analysis 4

2 Methodology and data 6

2.1 Background to sampling issues 6
2.2 The data 7

3 Theoretical model with random sample and potential biases 11

3.1 Random sampling 11
3.2 Proportionate sampling 11
3.3 Disproportionate sampling 12

4 Worked examples and conclusions 13

4.1 Random sample 13
4.2 Proportionate and disproportionate sampling 14
4.3 Possible mitigations and alternative approaches 16
4.4 Conclusions 18

5 Alternative application: estimating peak load 19

6 Conclusion 20

List of figures, tables and boxes

Matrix of sampling errors and coverage requirements (domestic only) 2
Table 2.1 Number of domestic and commercial properties 8
Table 2.2 Average annual consumption of domestic and commercial properties (kWh) 8
Table 2.3 Total annual consumption across all domestic and commercial properties (MWh) 8
Figure 2.1 Distribution of annual consumption across three settings ('000 kWh) 9
Table 2.4 Proportions of domestic properties in each setting 10
Table 2.5 Average annual consumption across groups and settings (KWh) 10

Table 2.6 Standard deviation of gas usage across groups and across settings (kWh) 10
Box 3.1 Sample size formula 11
Table 4.1 Coverage requirement calculations 13
Figure 4.1 Average percentage error of gas usage estimate across various coverage levels (%) 14
Table 4.2 Coverage requirements as a % of network across various sampling errors (%) 14
Table 4.3 Percentage of time sample accurately estimates network usage (%) 16
Table 4.4 Percentage of time sample accurately estimates the city network usage under disproportionate sampling (%) 16
Box 4.1 Water industry experience with smart meters and big data 17

Executive summary

The coverage requirement to estimate leakage with smart meters is high

This report provides a methodology and guidance for determining smart meter coverage requirements to estimate gas lost via leakage in the networks of the gas distribution networks (GDNs) in Great Britain.[1]

In principle, estimating leakage from a particular network will be possible by measuring (i) the total amount of gas put into the network and subtracting both (ii) the total amount of gas supplied to domestic customers and (iii) the total amount of gas supplied to commercial customers. Smart meters are currently being rolled out to domestic properties in Great Britain and, in this report, we consider whether smart meters could be used to estimate total domestic consumption. Oxera understands there are currently no plans in place to roll out smart meters to commercial properties, and we also comment on the implications of this for estimating total leakage.

The challenge in using smart meters to estimate total domestic consumption is that, until smart meters are rolled out to all properties in the networks, smart meter data will provide only a sample of usage data from these networks (we call the percentage of properties having smart meters the 'coverage'). For this reason, smart meter data cannot provide an exact value for total usage from all properties (and, by extension, of leakage), but only an estimate. We use the term 'sampling error' to provide a measure of how precise the estimate that results from the sample is as a percentage of total usage.

Since leakage is around 0.5–0.7% of gas usage, we consider the desirable sampling error should be around 0.1% (i.e. a sampling error tolerance of 10–20% of leakage), so that gas usage estimate based on smart meter data is sufficiently precise to assess leakage accurately. The domestic coverage requirements for smart meters across all three settings are very high when the required sampling error is 0.1% or less. A rural setting, in particular, would require smart meters in almost all properties in the network to achieve this level of precision. The city and town settings have lower coverage requirements but are still high (92–98% of network).

Moreover, in practice these estimates are conservative in two respects.

- The estimates are based on the domestic network only. Estimating total usage will require an understanding of consumption from commercial premises; Oxera understands there are no plans at present to install smart meters at these properties. Since the domestic and commercial networks are not separated, the high demand from commercial premises means that a sampling error of 0.1% of total usage would not be achievable using smart meter data alone, even with 100% coverage of domestic properties.

- The estimates are based on statistical theory that assumes random sampling. This is unlikely to be the case for smart meter rollouts. If some properties are more likely to have smart meters due to certain characteristics, gas usage estimates will be biased towards that particular group, and therefore will not give an accurate estimate of the whole network. If the bias is based on known and observed characteristics (i.e. data is available on these characteristics, such as age of property), it is possible to adjust for the bias. However, for any

[1] Although analysis is based on data for only one GDN—Wales and West Utilities.

unobservable sources of bias (for example, if smart meters were more likely to be installed in homes belonging to people of working age), such a correction will not be possible and the use of such data would risk giving a biased estimate of total usage.

It may be possible to reduce smart meter coverage requirements—for example, by gathering data from a point in time when domestic consumption is low and leakage is higher as a proportion of consumption (and thus easier to measure). However, this approach does not address the issue of the missing commercial data, and without access to actual smart meter data it is not possible to ascertain the extent to which such an approach would mitigate the high coverage requirements.

Statistical analysis therefore indicates that domestic smart meters are currently unlikely to offer a viable means of estimating total domestic consumption (and hence leakage) until they have reached a very high proportion of the domestic network.

Even at high domestic coverage, a separate method would be needed to estimate leakage from commercial properties, which will not be covered by smart meters in the short term.

Lower coverage is needed for lower precision estimates

There is a fundamental trade-off between the required precision and required smart meter coverage. Oxera considers that the 0.1% sampling error would be necessary to estimate leakage; however, if a higher sampling error were acceptable for certain situations, the coverage requirements would be lower across all three settings, as shown in the table below.

Matrix of sampling errors and coverage requirements (domestic only)

Acceptable sampling error (% of usage)	Error as a % of leakage[1]	% coverage		
		City	Town	Rural
0.1	17	92	98	100[2]
0.2	33	75	91	100[2]
0.4	67	43	72	99
0.6	100	25	54	97
0.8	133	16	40	94
1.0	167	11	30	92

Note: Based on a 90% confidence requirement, which is a standard approach in statistics.
[1] Oxera understands that actual leakage is around 0.5–0.7%; here it is assumed to be 0.6%.[2] Between 99.5% and 100.0%.

Source: Oxera analysis of Wales and West Utilities data

In this setting, it can be seen, for example, that around 11% coverage in the city setting is needed in order to obtain a usage estimate that is within 1% of the network average usage. However, this low level of precision is unlikely to be useful in this context since a 1% sampling error in total demand represents around 167% of leakage.

Other applications

Finally, smart meter data is potentially more promising in estimating peak load—i.e. '1 in 20 winter day' demand—than estimating leakage. This is because a less

restrictive sampling error is required, and daily smart meter data may be less varied than total annual gas consumption.

1 Introduction and overview

This report provides guidance on estimating smart meter coverage to assess gas leakage, and includes a discussion on factors determining coverage requirements. Three examples from various network settings illustrate how data variability and size of each network affect coverage requirements. We also discuss how these requirements change under various scenarios and potential application of smart meter data in estimating peak load.

1.1 Background

The gas distribution networks (GDNs) in Great Britain face regulatory incentives such as financial rewards (or penalties) according to performance against outputs. As part of the price control mechanism, GDNs have a set 'Shrinkage Allowance' for gas lost via leakage, in transportation or theft. In determining this allowance, GDNs are required to monitor and report network performance, produce a Shrinkage and Leakage Smart Metering Report on a periodic basis, and propose changes to improve the estimation of this allowance.

The GDNs have commissioned Oxera to provide advice on the use of sampling techniques to estimate network leakage using smart meters. Sampling theory suggests that a representative subset of data from smart meters could be used to estimate the aggregate (or population) usage, with some margin of 'sampling error'. This estimate could be compared with measured gas flows into the network to compute leakage. As smart meters are currently being rolled out to domestic properties only, our analysis focuses on the use of smart meters to estimate total domestic consumption. In the concluding sections, we comment on the lack of commercial smart meter coverage and what this implies in terms of estimating overall leakage.

1.2 Scope of analysis

This report provides guidance on choosing the right sample size, based on theoretical and practical design criteria. It includes three worked examples of the sample size calculation covering a range of gas distribution network settings, based on data provided by Wales and West Utilities (WWU). Specifically:

- a large city or local distribution zone (Bristol);
- a smaller town (Wrexham);
- a rural or more sparse network area (Bourton).

The report is structured as follows.

- Section 2 gives an overview of the methodology used to estimate sample size and describes the data available.
- Section 3 develops from the methodology introduced in section 2, presenting sample size formulae applicable to estimating coverage requirements. We also discuss approaches to sampling and their implications for estimation precision.
- Section 4 produces estimates of coverage requirements under the baseline case (sampling error of 0.1%) and shows how coverage requirements change across scenarios, using different sampling errors, a disproportionate sample, etc. We run simulations to confirm our results and discuss potential adjustments to address any sampling bias.

- Section 5 discusses how smart meter data might be applied to estimate peak load.
- Section 6 concludes.

2 Methodology and data

This section describes how sampling theory can be applied to estimate the smart meter coverage requirement in the GDNs. The section is technical in nature and some of the terminology used is as follows.

Population: all domestic properties in the relevant part of the network

Sample: properties with smart meters

Coverage: the proportion of all domestic properties with smart meters

2.1 Background to sampling issues

2.1.1 Sample size determinants

This report seeks to assess what coverage of smart meters is required to estimate leakage for a range of possible precision levels. The required coverage of smart meters—i.e. the sample size requirement—is determined by several factors, including the following.

- The higher the precision requirement, the larger the sample size needed. Oxera understands that leakage is around 0.5–0.7% of gas usage. The required precision in estimating leakage will depend on the context, but for the purposes of selecting a 'lower bound' for this analysis, Oxera assumes a minimum statistical precision requirement of 10–20% error around the true value.[2] A 0.1% error in estimating total gas usage is around 17% of leakage and therefore appears to be a reasonable benchmark for this purpose.
- The more varied gas usage is across a network, the larger the sample size that is needed. If gas usage varies considerably among properties, it is difficult to infer what network usage is based on a sample of smart meter data at a predetermined level of precision.
- Any potential bias towards selection of particular groups will also affect the sample size requirement. The standard sample size estimate is for a random sample—i.e. each property in a network is equally likely to obtain a smart meter and thus the sampling is random. However, if a particular group in the network is more likely to have smart meters then that group is more likely to be over-represented in the sample, and the proportion of that group in the sample will be much higher than its proportion in the network. In that case, the sample needs to be adjusted to provide an accurate reflection of network usage.

In addition, we make a technical adjustment using a 'finite population correction factor' if the coverage requirement is estimated to be over 5% of the network. When a sample takes up a large proportion of the network, it contains more information about the network than a sample with the same number of smart meters in a larger network. The finite population correction factor reduces the sample size required in this case. Intuitively, this captures the phenomenon that a small sample may be acceptable if it represents a very high proportion of the properties in the network.

[2] The precision requirement will depend on context, but 10% is frequently chosen as an upper benchmark. See, for example, Israel, G.D. (1992), 'Determining Sample Size', University of Florida, Florida Cooperative Extension Service, November, http://zulsidi.tripod.com/pdf/DeterminingSampleSizes.pdf, accessed 9 August 2016.

The precise formulae to be applied are shown in the following section.

2.1.2 Possible scenarios

As explained above, the sample size requirement will be determined by how the sample is constructed. While sample design is outside the scope of this report, we do consider three possibilities:

(i) the sample is random, which means that each property in the network is equally likely to obtain a smart meter;

(ii) the sample is proportionate, which means that the proportion of each group in the sample is equal to its size in the network (e.g. the sample will have the same proportion of houses with plastic gas pipes as the population as a whole);

(iii) the sample is biased towards/disproportionate in relation to particular groups whose consumption/leakage patterns may differ, and these groups are known (e.g. older houses with plastic pipes).

2.2 The data

2.2.1 Variables used in the analysis

Gas shippers are in the process of rolling out smart meters to their customers. However, smart meter data is not yet available as at the time of writing this report. Therefore, to estimate the smart meter coverage required, we base our analysis on the total annual consumption data provided by WWU.

The dataset includes commercial and domestic properties; however, it is Oxera's understanding that smart meters will be available for domestic properties only. As a result, it will be possible to estimate leakage only for domestic properties using smart meter data, and therefore we consider only these properties in our analysis.

To demonstrate the concept of sample size estimation, we set out three example calculations covering a range of GDN settings in WWU's networks:

- Bristol (a large city);
- Wrexham (a smaller town); and
- Bourton (a rural or more sparse network area).

Since each setting potentially has very different populations and patterns of gas usage, they may also have different smart meter coverage requirements. In section 4, we provide calculations for these three settings, taking into account these differences in estimating sample size requirements.[3]

As provided by WWU, the dataset used in this analysis includes:

- total annual consumption per property;
- age of property: before or after 1976;
- material of main pipe: plastic (i.e. polyethylene) or metal (i.e. cast iron, ductile iron, spun iron or steel).

[3] There may also be other differences, such as demographic ones, which we have not been able to control for. These are discussed in section 4.

The second and third variables are important explanatory factors since older properties, constructed under building regulations requiring less overall heat insulation, are likely to have higher gas usage (due to higher heating requirements), and older (metal) pipes are likely to have higher leakage.

2.2.2 Description of the data

Commercial versus domestic properties

WWU networks classify customers into domestic and commercial (i.e. shops and offices, schools and hospitals, hotels, pubs, clubs, restaurants and industrial). Since Oxera understands that, in the short term, smart meters will be installed only in domestic properties, it will not be possible to obtain gas usage data and estimate leakage for commercial properties with smart meters.[4] However, it is useful to check how the annual consumption of commercial properties differs from that of domestic ones, and what this implies for estimating leakage using domestic properties only.

Table 2.1 Number of domestic and commercial properties

	Commercial	Domestic	Total
City	4,646	173,979	178,625
Town	588	24,314	24,902
Rural	62	1,375	1,437

Source: Oxera analysis of WWU data.

Table 2.2 Average annual consumption of domestic and commercial properties (kWh)

	Commercial	Domestic	Total
City	149,223	12,059	15,626
Town	84,949	11,923	13,647
Rural	67,389	15,843	18,067

Source: Oxera analysis of WWU data.

Table 2.3 Total annual consumption across all domestic and commercial properties (MWh)

	Commercial	Domestic	Total	% Commercial consumption
City	693,289	2,097,926	2,791,230	25%
Town	49,950	289,891	339,840	16%
Rural	4,178	21,784	25,962	15%
Total	**747,416**	**2,409,601**	**3,157,032**	**24%**

Source: Oxera analysis of WWU data.

Commercial gas usage takes up a substantial proportion of total gas usage (25% in city and 15–16% in town and rural settings). Although the number of commercial properties across settings is low compared with domestic properties, the former have substantially higher gas usage (4–12 times more than domestic). Since the gas consumption pattern of commercial properties is

[4] In practice, only G4/U6-sized meters will be replaced by smart meters. They are the common domestic meter, which can cope with the average gas usage of a medium-sized home. Commercial properties are assumed to have larger meters and thus will not have smart meters installed in the short term.

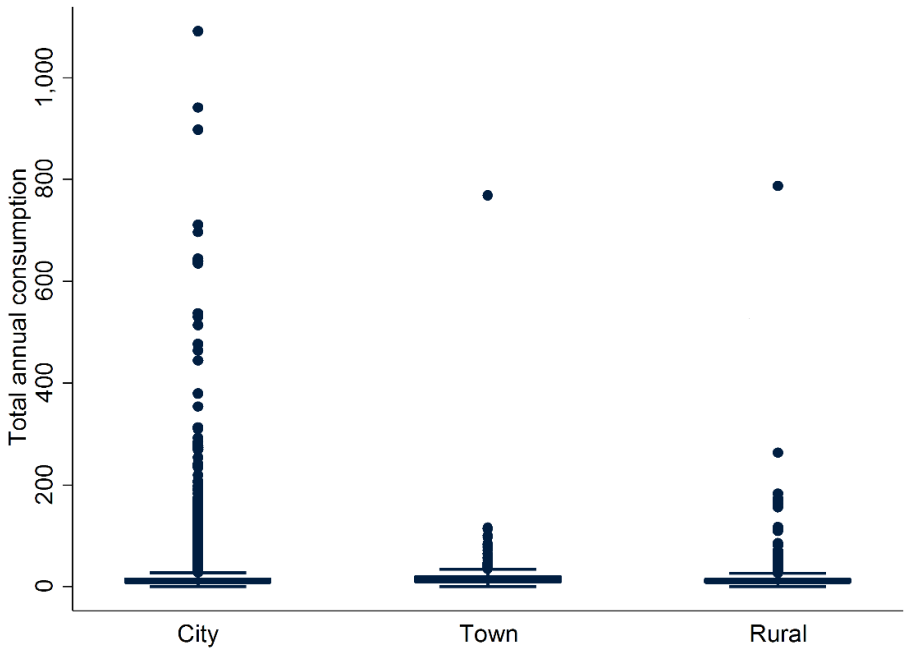
considerably different from that of domestic properties, any leakage estimate obtained from using smart meter data can only apply to domestic properties.

It is not possible to infer accurately, on the basis of domestic smart meter data alone, the gas usage of an entire network that includes a substantial proportion of commercial properties. In practice, Oxera understands that this finding will apply to most existing gas networks.

2.2.3 Removing extreme values from the dataset

The dataset contains observations with total annual consumption of 1kWh—unoccupied properties that we consider should be excluded from the analysis of how consumption varies between households. Moreover, data for town and rural settings contains observations with more than 600,000kWh, which is high compared with the main pattern of gas usage in these networks (see Figure 2.1). We remove these high values from the analysis in order to eliminate extreme variations in the data caused by only two observations in the town and rural settings.

Figure 2.1 Distribution of annual consumption across three settings ('000 kWh)



Source: Oxera analysis of WWU data.

2.2.4 Breakdown of the dataset into four groups

Below is the breakdown of proportions based on meter age and main pipe material, after removing the extreme values, as discussed above. The largest group is pre-1976 properties with plastic main pipe, while the smallest group is post-1976 properties with metal main pipe, especially in the city setting.

However, the city setting has more than 10 times the number of observations than the rural setting, which will influence the sample size requirements of these two areas.

Table 2.4 Proportions of domestic properties in each setting

	Post-76 and plastic pipe	Post-76 and metal pipe	Pre-76 and plastic pipe	Pre-76 and metal pipe	Total no. of observations
City	13%	2%	69%	16%	173,979
Town	23%	8%	47%	22%	24,314
Rural	12%	4%	59%	25%	1,375

Source: Oxera analysis of WWU data.

On average, properties in the rural settings use the most amount of gas, while those in the city use the least.

Table 2.5 Average annual consumption across groups and settings (KWh)

	Post-76 and plastic pipe	Post-76 and metal pipe	Pre-76 and plastic pipe	Pre-76 and metal pipe	Total no. of observations
City	10,104	10,230	12,198	13,219	12,059
Town	12,362	13,643	11,103	12,596	11,923
Rural	15,291	16,068	15,164	17,715	15,843

Source: Oxera analysis of WWU data.

In addition, variability of the data on gas usage, as measured by standard deviation, is highest in the rural setting.[5] Below are the standard deviations for each group of properties and for each network in total.

Table 2.6 Standard deviation of gas usage across groups and across settings (kWh)

	Post-76 and plastic pipe	Post-76 and metal pipe	Pre-76 and plastic pipe	Pre-76 and metal pipe	Total no. of observations
City	18,626	13,902	8,478	9,748	10,619
Town	8,422	8,318	6,663	6,835	7,281
Rural	14,503	12,567	11,253	11,404	11,793

Source: Oxera analysis of WWU data.

[5] This is approximately the average difference between each property and the average usage level.

3 Theoretical model with random sample and potential biases

3.1 Random sampling

We estimate sample size using variability of all domestic properties in each network. As discussed above, the assumption here is that each domestic property has an equal chance of obtaining a smart meter and falling into our sample.

The inputs for this calculation are:

- the average level and standard deviation of annual gas consumption for all properties in each network;
- a maximum sampling error allowance of 0.1%, which measures the accuracy of our estimate (see section 2.1);
- a confidence level requirement of 90%. This specifies that if we take a sample from a network, we can expect that gas usage of the whole network is, on average, within 0.1% of usage by the sample as a whole for 90% of the time. This assumption enters the sample size calculation in the form of a z-score of 1.645 for the 90% confidence level.

Statistical theory allows these factors to be combined to generate a sample size requirement. The relevant formulae assuming random sample are shown below.

Box 3.1 Sample size formula

The sample size formula is as follows:[1]

sample size = [(1.645 x standard deviation)/(average x 0.1%)] [2]

If the sample size needed is more than 5% of the network, a finite population correction factor is applied:[2]

adjusted sample size = sample size/[1 + (sample size – 1)/network size]

Source: [1] Berenson, M., Levine D.M. and Krehbiel, T.C. (2005), Basic Business Statistics, Pearson, http://courses.wcupa.edu/rbove/Berenson/10th%20ed%20CD-ROM%20topics/section8_7.pdf, accessed 9 August 2016. [2] Israel, G.D. (1992), 'Determining Sample Size', University of Florida, Florida Cooperative Extension Service, November, p. 4, http://zulsidi.tripod.com/pdf/DeterminingSampleSizes.pdf, accessed 9 August 2016.

3.2 Proportionate sampling

With proportionate sampling, the size of each group (based on meter age and pipe material) in the sample is proportional to its size in the network. For example, if pre-1976 properties with plastic pipe represents 70% of the city network, this method assumes that it is also 70% in the sample for that network. Proportionate sampling can increase estimation accuracy if the variability of data *within* each group is small. Thus, when we group properties with similar gas usage and install smart meters in some of them, smart meter data from each group is a good representation of the gas usage of the whole group.[6]

We therefore compare random and proportionate sampling using simulation to check if proportionate sampling improves estimation accuracy in this case.

[6] Kish, L. (1965), *Survey Sampling*, John Wiley & Sons, p. 76.

Simulations are run as follows. We draw 1,000 samples for each method, and then estimate and compare the number of times the gas usage of the whole network falls within 0.1% of the gas usage of the sample drawn.

3.3 Disproportionate sampling

The two methods discussed above assume that each domestic property has an equal chance of obtaining a smart meter and falling into our sample. However, smart meter installation is not a random process, but is determined by customers' level of interest in having a smart meter installed in their homes. Therefore, smart meter data, especially at the beginning, will be available only from properties whose owners register their interest in having a smart meter.

In this case, in a practical setting we may expect certain groups of customer to be disproportionately sampled.[7] These groups will take up a larger proportion of the sample than their size in the network. To address this, we adjust the gas usage estimate obtained from this type of sample by weighting the average usage of each group by its size in the network. The adjustment brings the gas usage estimate closer to the network gas usage and increases accuracy.[8]

In section 4.2, we consider the case where the group of older properties with plastic pipes is more likely to have smart meters, and thus more likely to be in the sample. This provides an illustration of the largest potential impact of disproportionate sampling since this group has the lowest variation in gas usage among the four groups in our data (in general, an efficient sample should be targeted to the highest-variance group). Disproportionate sampling of this group, means that the sample is unlikely to reflect the true variations in gas usage in other groups of the network.

[7] There has not been enough data to explore and identify patterns of customers getting smart meters. Oxera understands that, to date, the demand for smart meters seems sporadic, which may mean that random and proportional sampling methods are applicable.
[8] Kish, L. (1965), *Survey Sampling*, John Wiley & Sons, p. 90.

4 Worked examples and conclusions

4.1 Random sample

4.1.1 Coverage requirement calculations

Using the formulae explained above, we obtain coverage requirements for all three settings.

Table 4.1 Coverage requirement calculations

	Average	Standard deviation	Unadjusted coverage	Network size	Adjusted coverage	Percentage of network
City	12,059	10,656	2,113,046	173,979	160,744	92%
Town	11,923	7,329	1,022,510	24,314	23,749	98%
Rural	15,843	11,828	1,508,129	1,375	1,374	100%

Source: Oxera analysis of WWU data.

4.1.2 Coverage requirements across various sampling error conditions

As described above, the coverage requirements here are calculated based on the desirable precision level—i.e. sampling error—of 0.1% of total use, equivalent to around 10–20% of leakage. This means that if we have smart meter data according to these coverage requirements, we can be reasonably sure (with 90% confidence) that the gas usage for the whole network is within 0.1% of the gas usage estimate obtained from smart meters.[9]

As we reduce the coverage of smart meters used, the estimate will be less accurate, leading to a significant risk of unacceptably high sampling errors (larger than 0.1%).

Average percentage error

For each setting we simulate 1,000 samples from the annual consumption data and calculate the absolute percentage difference between gas usage from each sample and from the whole network. The average percentage error over these 1,000 samples for each setting across various coverage levels is shown in Figure 4.1 below. For example, if smart meter coverage in a city is 20%, we can expect that, on average, the gas usage estimate from smart meter data is 0.35% away from the gas usage of the whole network. On the other hand, if coverage is 80% in a small town, smart meter data can, on average, provide a gas usage estimate within 0.16% of the whole network usage.

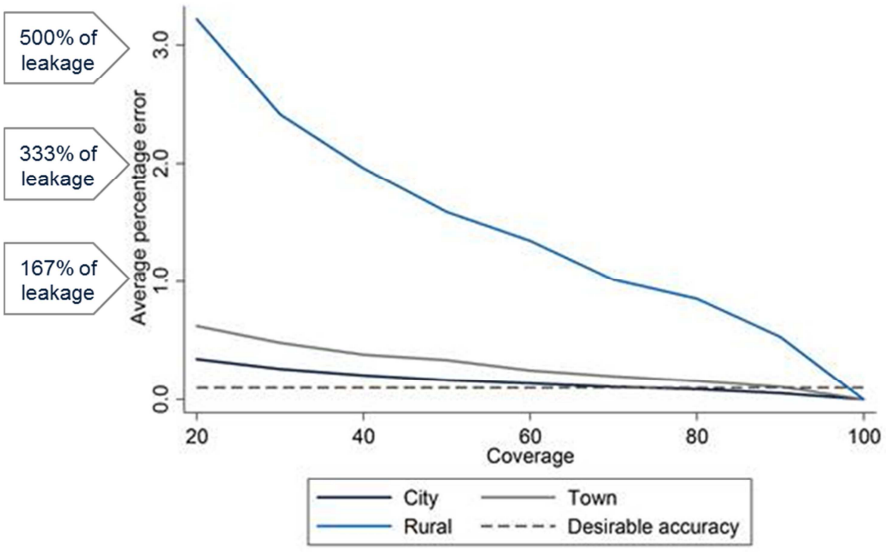
The calculation in this section (see Figure 4.1) is slightly different from the coverage requirement calculation in Table 4.1.

- Figure 4.1 shows that, to achieve an average absolute error of 0.1% across 1,000 samples, we need smart meter coverage of 70–80% in a city.
- In contrast, the coverage requirement calculation does not take into account the magnitude of sampling errors, except whether the errors are lower or higher than 0.1%. It computes the coverage needed to have a gas usage estimate within 0.1% of the network usage 90% of the time (our confidence

[9] Some caution is needed in interpreting the 90% confidence level. Technically, it is not directly equivalent to being 90% sure that the 'true' (population) usage is within the sample estimate, although the distinction is not crucial in this report. It is a standard threshold for a reasonable degree of confidence in applied statistics.

level requirement). The coverage requirement for a city is 92% (see Table 4.1).

Figure 4.1 Average percentage error of gas usage estimate across various coverage levels (%)



Note: Errors expressed as percentage of leakage assume that leakage is 0.6%.

Source: Oxera analysis of WWU data.

Varying sampling errors

To illustrate how sampling error affects coverage requirements, we consider scenarios where sampling error varies from 0.1% to 1%. As we relax the accuracy requirement by increasing the sampling error, the coverage requirement decreases across all three settings (see Table 4.2).

Table 4.2 Coverage requirements as a % of network across various sampling errors (%)

Sampling error (%)	Error as a % of leakage[1]	City	Town	Rural
0.1	17	92	98	100
0.2	33	75	91	100
0.4	67	43	72	99
0.6	100	25	54	97
0.8	133	16	40	94
1.0	167	11	30	92

Note: [1] Oxera understands that actual leakage is around 0.5–0.7%; here it is assumed to be 0.6%.

Source: Oxera analysis of WWU data.

4.2 Proportionate and disproportionate sampling

4.2.1 Simulation methodology

Section 3 defined two methods of sampling: proportionate and disproportionate sampling. Simulation of these methods allows us to calculate how well each

performs in providing an accurate estimation of network usage, when compared with random sampling.

The simulation approach is undertaken using the following steps.

- Choose from the dataset for each setting a number of observations (annual consumption in this case) based on Table 4.1: city 92%, town 98% and rural ~100%.

  i) for 'random sampling', the selection is purely random;

  ii) for 'proportionate sampling', the selection is random within each group, but we ensure that the proportion of each group in the sample is the same as its size in the network;

  iii) 'disproportionate sampling' assumes that 100% of older properties with plastic pipes have smart meters, and thus belong to the sample, while properties in other groups are chosen randomly. This approach represents an extreme form of selection bias and therefore an indication of the maximum impact of this bias.[10]

- Check if network usage is within 0.1% of the estimate obtained from the sample drawn above.

- Repeat the first two steps 1,000 times and count the number of times that network usage is within 0.1% of the smart meter estimate.

4.2.2 Simulation results

The simulation results are presented in Table 4.3 below, which shows the percentage of times (out of 1,000 repetitions) that each method accurately estimates network usage—i.e. network usage is within 0.1% of sample estimate.

- Since the confidence level used is 90%, the random sampling method, as expected, accurately estimates network usage around 90% of the time across all three settings.

- Proportionate sampling, in this case, does not seem to improve estimation accuracy due to high variability of gas usage within each group of properties. (This approach improves accuracy if variability is low within each group.)

- Disproportionate sampling, on the other hand, negatively affects the estimation accuracy level. Out of 1,000 samples drawn with this assumption, the percentage of times the sample accurately estimates network usage reduces significantly, especially in the city setting. As highlighted before, we need to weight the average usage for each group by its size in the network. After this adjustment, the accuracy level is considerably higher than before and close to that of the random and proportionate samples.

[10] As discussed above, rollout of smart meters is based on customers' interest. As such, smart meter data, especially at the beginning, may not provide a random sample. Disproportionate sampling may therefore be applicable in this case. The group of older properties with plastic pipes is chosen to illustrate the application of disproportionate sampling since it has the largest proportion in the networks and the lowest variation in gas usage.

Table 4.3 Percentage of time sample accurately estimates network usage (%)

Setting	Random sample	Proportionate sample	Disproportionate sample	Adjusted disproportionate sample
City	90	89	7	72
Town	91	90	24	88
Rural	90	89	88[1]	88[1]

Note: [1] The high accuracy level of disproportionate sample in the rural setting is due to the high coverage requirement of almost 100%. Therefore, sampling methods do not affect the results significantly.

Source: Oxera analysis of WWU data.

The bias, under disproportionate sampling, towards older properties with plastic pipes has the worst effect on estimation precision in the city setting of all the sampling approaches. After adjustment, we obtain the desirable sampling error of 0.1% only 72% of the times, using samples of 92% of the city network. Thus, to obtain the sampling error of 0.1% for around 90% of the times, the coverage for city should be higher than 92% under disproportionate sampling.

The results from simulations across various coverage levels in the city setting are presented in Table 4.4 below. It shows that smart meter coverage of up to 96% would be necessary to obtain the sampling error of 0.1% for 90% of the times if a bias towards older properties with plastic pipes exists in the sample.

Table 4.4 Percentage of time sample accurately estimates the city network usage under disproportionate sampling (%)

Coverage (%)	Disproportionate sample	Adjusted disproportionate sample
92	7	72
94	16	83
96	42	91
98	88	98
100	100	100

Source: Oxera analysis of WWU data.

This method of adjustment also applies to other biases given that we have data on the characteristics causing these biases. For example, if properties with owners of working age are more likely to have smart meters and lower consumption, this bias may translate into large estimation error since smart meter data available is not representative of the whole network. Oxera understands that WWU currently does not have access to information on customer demographics. If such data becomes available through the take-up of smart meters, this type of bias can be corrected in a similar way as we demonstrate here with bias due to meter age and pipe material.

4.3 Possible mitigations and alternative approaches

The results shown so far indicate that with the desirable sampling error of 0.1%, coverage requirements across network settings are high (over 90% in all cases). That means there are significant challenges in applying smart meter data for the purpose of leakage estimation, at least in the short run when coverage is still relatively low.

However, in the longer term it may be possible to apply more complex methodologies to use smart meter data to obtain a more accurate estimate of

total usage. Actual smart meter data would be required to assess the viability of these techniques.

First, if smart meter data is collected when gas usage is expected to be very low (for example, at night on the warmest day of the year), it is feasible that a higher proportion of the gas flowing into the network will relate to leakage.[11] This method of collecting smart meter data affects two variables in our coverage requirement calculation:

- we can allow for higher sampling error as leakage is expected to represent a larger proportion of gas usage; and
- low to almost zero gas usage across properties during that specific window means lower variability in smart meter data, which also lowers the coverage requirement.

Moreover, with smart meters, gas usage data for each property may be collected at multiple points in time during a year, creating a richer dataset than currently available without high smart meter coverage. Smart meters allow access to gas usage across an almost constant group of properties (those with smart meters) over a period of time (a panel dataset), which can deliver higher estimation precisions than a dataset of the same number of properties at only one point in time (a cross-sectional dataset).

More broadly, when smart meter data is available, techniques in big data can be employed to increase precision. Instead of predicting which customer characteristics are most beneficial in determining gas usage, with access to smart meter data, we can group properties with similar usage variability together, which may allow for more precise estimation (although the coverage requirement will again be high). Some insight on water industry experience with big data is provided in Box 4.1.

Box 4.1 Water industry experience with smart meters and big data

The water sector has considerable experience in deploying and applying analytical techniques to big data derived from smart meters. An early study by the Wide Bay Water Corporation in Queensland, Australia, used relatively wide interval data (hourly flow) to characterise customer consumption. This allowed consumption to be studied at a more granular level, and for leakage flows to be separated from normal water consumption. Further studies have since allowed more granular disaggregation of usage into specific end uses, such as showering or the use of washing machines.

In terms of leakage assessment, these deployments have tended to focus on post-meter or within premises leakage, which is a water sector-specific policy consideration. Nonetheless, the technique of using night-time flow to estimate leakage post-meter could also be applied to the problem of network-side leakage. This works by defining aggregate flows at the time of lowest demand as leakage. Currently in Great Britain, GDNs do not have access to a daily or diurnal gas flow measure, and so this approach is difficult.[12] If such a detailed network flow measure were available (for example, by using flow meters installed at upstream network locations), this approach could be applied in the

[11] This assumes leakage at a given point in time is unrelated to the level of consumption happening at that point in time. However, commercial loads may not follow the traditional diurnal profile, adding significant uncertainty. Moreover, measuring gas input at a specific date and time is not currently possible, as WWU does not have local district metering.
[12] The application of using night-time flow in gas leakage estimation is discussed in section 4.3.

gas setting. This would then require a granular end-user consumption measure for use as a comparison.

Researchers and utility operators in the water and electricity sector have developed methods for profiling consumption habits using big data techniques and finely detailed, spatio-temporal meter readings. These algorithms are similar to a type of cluster analysis that is capable of automatically grouping consumers by similar usage 'signatures'. This can also segment consumption along contextual lines, such as weather, season, location or vacations. Other techniques such as principle component analysis (PCA) and time-series regression analysis can be combined with customer profiling to estimate total demand. By doing this in a data-led, automated way (rather than relying on other demand covariates, such as house type), more efficient sampling techniques can be devised for demand estimation.

These big data techniques have been developed after a substantial amount of water smart meter data had become available. It is not possible to tell yet whether and to what extent similar techniques can be applied in the gas industry.

Source: Britton, T., Cole, G., Stewart, R. and Wiskar, D. (2008), 'Remote diagnosis of leakage in residential households', *Water*, September. Willis, R. Stewart, R.A., Panuwatwanich, K., Capati, B. and Giurco, D. (2009), 'Gold Coast domestic water end use study', *Water*, September. Alahakoon, D. and Yu, X. (2013), 'Advanced analytics for harnessing the power of smart meter big data', Conference paper, IEEE International workshop on Intelligent Energy Systems, November. Wijaya, T.K., Ganu, T. Chakraborty, D., Aberer, K. and Seetharam, D.P. (2014), 'Consumer Segmentation and Knowledge Extraction from Smart Meter and Survey Data', Conference paper, SIAM International Conference on Data Mining, April.

4.4 Conclusions

Overall, leakage estimation requires high smart meter coverage across all three settings if we consider the necessary sampling error at 0.1%. The rural area, in particular, needs almost 100% coverage to deliver this level of precision. This is due to a smaller number of properties and relatively high variability in annual consumption data in the rural setting. Moreover, this study has considered domestic properties only; allowing for the additional uncertainty that arises due to smart meters not covering commercial properties, the resulting precision may be even lower than we have allowed for.

Once a significant body of smart meter data is available, alternative techniques may be applied to increase the precision of smart meter-based total usage estimates. However, it is not possible to assess the viability of these methods prior to the establishment of such a dataset.

Statistical analysis therefore indicates that domestic smart meters are currently unlikely to offer a viable means of estimating total domestic consumption (and hence leakage) until they have reached a very high proportion of the domestic network.

Even at high domestic coverage, a separate method would be needed to estimate leakage from commercial properties, which will not be covered by smart meters in the short term.[13]

[13] This conclusion is based on the sample provide to Oxera. Whether these results apply to other networks (including those within WWU besides Bristol, Wrexham and Bourton in the analysis, or other GDNs) depends on how different the data variability is across networks. If gas usage variability is similar in networks of the same setting, we can expect the coverage requirements calculated in this analysis to apply to other networks. Otherwise, the same methodology proposed here can be used to estimate coverage requirements of specific networks.

5 Alternative application: estimating peak load

GDNs are required to have sufficient capacity to meet peak winter gas demand, defined as a '1 in 20 winter day'.[14] Oxera considers that daily smart meter data collected over several years is likely to be useful in estimating this peak load.

With smart meter data, we can estimate the relationship between gas usage and weather conditions (such as relative temperature and wind speed). Simulations of gas usage can be run based on this relationship, which are then fitted to a statistical distribution.[15] The peak load—i.e. 1 in 20 demand—is estimated at around the 95% value of this distribution.

Smart meter data can also be used in estimating the diversity factor—in this case, the ratio between the sum of maximum gas usage across all properties in the network and the true maximum gas usage of the network. In reality, not all households have their maximum usage at the same time. Therefore, summing peak loads across the whole network is likely to overestimate the network's true peak load. Daily smart meter data collected over time will allow GDNs to estimate the diversity factor over several years and adjust for this overestimation of the network peak load.

Peak load estimation does not have the same restrictive condition on sampling error of 0.1% as with leakage estimation. As a result, it will not require as high smart meter coverage. However, similar to leakage estimation, more precise peak load estimate will require higher smart meter coverage. Most importantly, because the statistical distribution can be inferred from a sample of daily data, it will not be necessary to observe 20 winters to estimate the '1 in 20 winter' peak demand. This could instead be derived from fewer years of data.

Therefore, in the short term at least, smart meter data is more promising as a method for estimating peak load than it is for estimating leakage.

[14] The level of demand that, in a long series of winters, with connected load held at the levels appropriate to the winter in question, would be exceeded in one out of 20 winters, with each winter counted only once. See www2.nationalgrid.com/WorkArea/DownloadAsset.aspx?id=36706, accessed 29 May 2016.
[15] Technically, a lognormal, Weibull or Gumbel–Jenkinson distribution. See National Grid (2012), 'Gas Demand Forecasting Methodology', February, p. 37.

6 Conclusion

This report has provided a thorough discussion on sample size determinants that are applicable to calculating smart meter coverage requirements. They include the level of data variability, desirable sampling error and confidence level requirement. We also consider other factors important in sampling, such as any potential biases in a sample and how to adjust the estimate to increase precision in that case.

We find that with the baseline conditions (sampling error of 0.1%, which is the desirable minimum precision in leakage estimation, representing a margin of error of 10–20% of total leakage), smart meter coverage requirements are high across all three settings in the analysis (city–Bristol, town–Wrexham and rural–Bourton). These requirements vary from 92% (city setting) to almost the entire network (rural setting). Simulations of three sampling techniques (random, proportionate and disproportionate sampling) confirm our findings and compare the precision level of these techniques.

Moreover, this study has considered domestic properties only since commercial properties will not initially be covered by smart meters. An alternative method will therefore be needed to estimate commercial consumption. Allowing for the additional uncertainty that arises due to smart meters not covering commercial properties, the resulting precision will be even lower than we have allowed for.

Once a significant body of smart meter data is available, alternative techniques may be applied to increase the precision of smart meter-based total usage estimates. However, it is not possible to assess the viability of these methods prior to the establishment of such a dataset.

We conclude that if smart meter data is to be applied effectively to leakage estimation, a very high proportion of the domestic network will need to be covered by smart meters and an alternative source would be needed to estimate commercial usage. However, we consider that lower smart meter coverage than those computed in this analysis may be useful in estimating peak load for domestic properties.

www.oxera.com